DOI:
10.1145/2470654.2466174

Document Version
Publisher's PDF, also known as Version of record

Putting Things in Focus: Establishing Co-Orientation Through Video in Context

James Norris, Holger Schnädelbach
Mixed Reality Laboratory
University of Nottingham,
Nottingham, UK
{psxjn, holger.schnadelbach}@nottingham.ac.uk

Paul Luff
Work Interactiona and
Technology Research Centre
King's College London, UK
Paul.Luff@kcl.ac.uk

ABSTRACT
In collaborative video communication systems, establishing co-orientation around physical objects, virtual objects and people is a critical requirement. This is problematic as the technical limitations of video fractures the display of coduct in the connected environments. We present the results of a study of one collaborative system, CamBlend, which aims to alleviate some of these problems by using screen based pointing tools to both physical spaces and virtual resources. We report on how participants achieved co-orientation when using this system. We relate these findings to previous research into the fractured ecologies of collaborative spaces, describing how the form and nature of fractures in CamBlend differ from earlier reported work.

Author Keywords
CSCW; collaboration; interaction analysis; focus+context.

ACM Classification Keywords
H.4.3. Communications Applications: Computer conferencing, teleconferencing, and videoconferencing

INTRODUCTION
Co-orientation is the capability of multiple communicating parties to establish joint attention around a common focal point. Successful co-orientation is a foundation for many collaborative processes such as turn taking in group meetings [23], utilization of physical resources (e.g. whiteboards, printed documents) [15, 18] or digital resources (e.g. presentations, digital documents). In face-to-face communication, we use a range of mechanisms to establish co-orientation. These mechanisms can be explicit, spoken directives, e.g. 'look at that'. However, normally co-orientation is achieved with minimal visible effort through a range and combination of subtle physical cues such as turning the head, glancing or gesturing. This is sometimes referred to as 'projection' [24]. We take the definition of projection from Kuzuoka et al. [14] as: *'... the capacity of participants to predict, anticipate, or prefigure the unfolding of action.*'. The process of pointing and directing attention through gesture is itself a complex and collaborative process [11]. The trajectory of the pointing motion guides attention as it moves, both in its production to grab attention through to completion in order to guide that attention towards an object. Hindmarsh & Heath [11] also discuss the complex relationship between pointing action and the deictic reference (a directive that requires context to be understood, e.g. 'this', 'that') that might support them.

When collaborating at a distance, video seems particularly valuable, with desktop video collaboration now common within the workplace. Existing video communication tools make projection and therefore co-orientation difficult. Largely due to technical limitations of the pin-hole camera model, communicating through video introduce a range of interactional difficulties, particularly when referring to objects. In this paper we focus on the problem of co-orientation, an issue addressed by a number of novel systems and contexts [13, 14, 15, 16, 19, 21].

The problems which limit the capacity for co-orientation have been described in terms of the 'dual ecologies' created when connecting two remote spaces [14]. For example Gaver [6] discusses how the affordance of predicable interaction and the anisotropy of video media results in a discontinuous space for action. Heath et al. [9] discuss how activities (including the abilities to project action) are fractured when communicating through video, and suggest how the fractured ecology of action has consequences for participants' abilities to ground their actions in relation to the objects in the local and remote environments [15].

In this paper we consider how one particular solution, using a way of presenting both focus and context, can support co-orientation to objects in a video-mediated environment. This seems to constribute to how we understand the nature of action in mdiated spaces and the way activties are fractured or fragmented

CO-ORIENTATION TOWARDS OBJECTS AND PEOPLE
The problems of fractured ecologies within video communication mean that the projective cues necessary for co-orientation seem less effective when mdiayed through video than when face-to-face. People then appear to adapt their turn-taking social behavior when using video. A number of studies have focused on the adaptability of

CHI 2013, April 27–May 2, 2013, Paris, France.

people and the methods used to achieve co-orientation with people over video. It is shown that people attempt to compensate for these failures by adopting a more formal, explicit form of communication. For example O'Conaill & Whittaker [20] found that low-quality video communication contained fewer backchannel markers, more explicit handovers and a more explicit form of turn-taking. Similarly, Martin et al. [17] showed there to be extensive extra work required in video banking consultations to manage the customer, consisting principally of 'face' work, including exaggerated smiling, facial gestures and nodding. Further evidence of the capacity of users to adapt their behavior to cameras comes from Aanested [2], from studies on video mediated surgery, showing the ability of surgeons to 'perform' to a camera. These studies also highlight the capacity of users to modify their behavior to accommodate a technological moderator.

Orientation around physical objects often has a practical requirement, such as when the focus of the discussion is the object (e.g. when architecture studios collaboratively discuss a prototype, the model is the focus of attention). This requirement is the focus of much research into 'expert helper' systems, where detailed collaboration around a single physical object is required [4]. It is often argued that video has the most potential, or utility when used on object orientated tasks like expert-helper tasks, rather than person focused tasks [18]. However, Luff et al. [15] show the general utility of the local environment even outside of object focused tasks. They show that many normal collaborative actions are embedded in features of the local environment, which engenders particular action and serves as a "center of coordination".

As well as the need to ground our action in physical objects, there is an increasing practical requirement to coordinate action around virtual resources, which often integrate, or form the contextual backdrop to meetings (e.g. reviewing documents). These resources can be on separate devices, for example physical laptops located on one side of the collaboration, while this causes asymmetrical access and is subject to the same limitations as orientation around physical objects. Most commercial desktop utilities such as Skype [1] recognize the utility of inline integration of shared virtual resources by including features such as document or desktop sharing. There has been renewed interest in more advanced forms of integration, combining physical and virtual referencing in high-end (i.e. expensive and custom-built) 'blended' media spaces[16, 21]. Here the focus has been to maintain spatial consistency by careful calibration of cameras. These systems seem successful in presenting gaze and bodily orientation coherently. However, in systems like Halo and BISi, providing a distinct display to present virtual and physical objects makes assessing another's orientation to these problematic [21]. A problem partly resolved in systems like tRoom where displays of objects are integrated into those of co-participants, and yet this requires only particular kinds of ways of working [16]. All these blended systems involve different trade-offs with respect to system cost, portability and flexibility. They also restrict the environment in which actions take place.

CAMBLEND

In this paper we explore co-orientation through a team based collaborative design task, using an enhanced version of CamBlend [19]. CamBlend is based on a focus + context video design where the manipulation of the focus windows provides a method for directing and interpreting attention. The utility of focus + context video for collaboration has been explored previously through a number of similar systems [3, 25]. In what follows, we present a brief system overview of CamBlend focusing on the changes that have been made to the original system.

System Summary

CamBlend is a bi-directional panoramic conferencing system which provides high resolution (3072x768) access to local and remote spaces. Each space is represented on-screen separately in a video frame, shown in figure 1. Both frames have identical features, users therefore have symmetrical access to their local and remote environments.

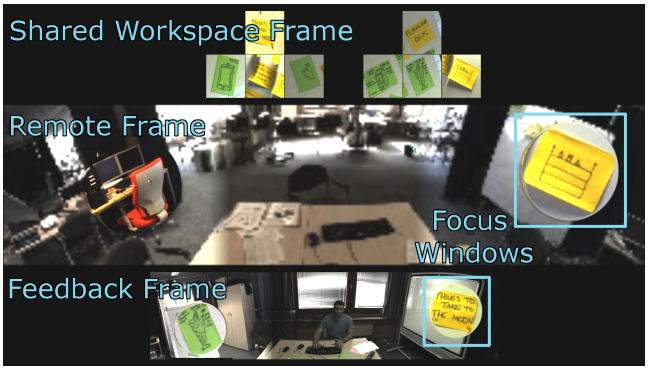


Figure 1. A CamBlend screen, showing the different frames

Each frame contains two high-resolution focus windows, which can be manipulated to show specific portions of the physical space in detail. These focus windows are manipulated over a low resolution contextual background which shows 180° field of view (see also figure 2). Manipulation of these windows is visible to both sides, so moving a window in the remote frame (i.e. which shows the remote side) can be interpreted by the remote side as stating 'I am now looking here'. Equally manipulation of local windows equates to 'I want you to look here'. Teams of participants have equal access to these resources via two mice on each end of the system, making a total of four inputs (although there is no fixed limit). These mouse pointers each have symmetrical access to manipulate any of the four available focus windows.

Shared Workspace

In order to study the co-orientation around screen based virtual resources, a range of tools for collaboratively organizing and referencing virtual resources inside a shared workspace have been implemented. The requirement for

these features also ties into previous studies of CamBlend, which highlighted the need for more advanced handling of virtual resources [19].

This shared workspace area is located at the top of the screen, shown in figure 1. The workspace functions as a grid of resources, which can be seen, retrieved, deleted and organized symmetrically by both sides of the interaction. The grid operates over 20 shared slots arranged on a grid of 2x10. Each slot can contain a 'snapshot' of the contents of one of the focus windows, i.e. a snapshot of a physical resource. Users have the ability to move these snapshots around the grid, creating relationships or categorizations between them. Both terminals have synchronous, symmetrical access to manipulate this frame. Users also have the ability to click on any occupied slot to enlarge the contents. This is visible to both sides as a mechanism to 'highlight' or indicate a specific snapshot. This function is intended to support deictic referencing to the item of interest, i.e. by supporting a user in their ability to click on the snapshot and declare their interest in 'this one'.

As the brief description highlights, CamBlend falls into an existing category of proposed solutions that provide pointing tools and some combination of gestural or directing tools [5, 19]. These systems show promise and have the primary benefit of being lightweight, usually requiring little specialist equipment and a very flexible deployment environment. Similar to GestureMan [15], these systems provide tools which aim to replace physical gestural and projective actions (i.e. by controlling a laser pointer instead of pointing), but in CamBlend we do not require an additional physical device to support referencing.

STUDYING CO-ORIENTATION IN CAMBLEND

A lab based quasi-naturalistic user study was undertaken using CamBlend. Our primary objective was to gain insights into our participants' practical accomplishment of co-orientation around objects, both local or remote, virtual or physical, as well as orientation around people. The study design was for teams of engineering students to creatively design and develop a product. Practically, participants were guided through a number of phases that would result in a final deliverable. These task phases were designed around the widely used creative problem solving process (CPS) [22]. The foundation of this process is in the expand and refine principle. In principal, to begin with lots of diverse ideas are generated individually and then collaboratively teams worked to structure and eliminate those ideas. Finally the selected idea is refined until ready for delivery.

Figure 2. The physical CamBlend setup

Task Overview

Participants were asked to design two related items, a perfume bottle and the storefront stand that it should display on. Participants were provided with a short brief which described the type of scent and its target market. They were asked to complete this design exercise in teams of 4. Engineering was chosen as a discipline which often works with physical or virtual objects, including architectural models, product prototypes or building materials. Participants were recruited from the local university and ranged through all years of undergraduate engineering degree, to Masters and PhD level students. There were a total of 28 participants split into groups of 4, making 7 teams in total. All teams were reimbursed at a rate of £10 per hour. Session lengths varied but usually lasted for 90 minutes. Participants were also offered a prize of £60 per team for the best design. Members from five of the teams had met before while two teams were matched by the moderator. All participants had normal, or corrected-to-normal vision. In order to reflect the requirements of CPS [22] the task was split into five sections:

Idea Generation Individually brainstorm perfume bottle ideas (5 minimum) on large post-its. Participants worked alone on generating ideas.

Idea Presentation Present ideas to the group. Each team member in turn presented their ideas back to the group, highlighting the reasons for their design.

Idea Clustering Collaboratively identify the relationships between the ideas generated and generate a categorization structure to organize the ideas. Then arrange those ideas into the devised categories.

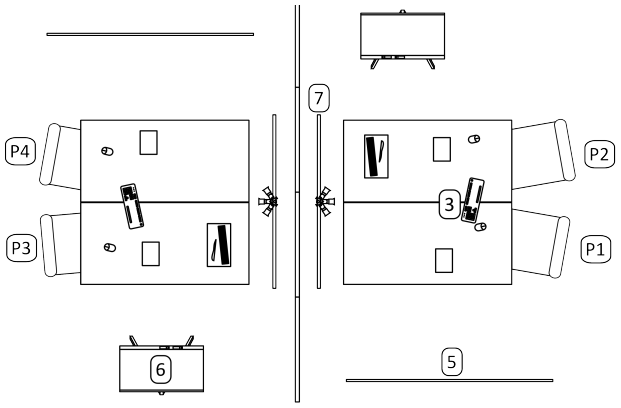
Idea Selection Use the cluster map to agree on a single idea, or a combination of ideas to move forward for final development. Participants were encouraged to use whatever collaborative mechanism they wanted to decide on an idea.

Idea Development Develop a final deliverable version of the product. Participants were provided with a range of resources and were free to choose whatever they felt would best represent their design.

Participant Instructions

Participants were taken through the steps described in turn but they completed the stages in their own time. The

moderator stood at the divide to communicate with all 4 members of the team simultaneously and to provide the stage instructions as the teams moved between them. The instructions were provided as described above, with the exception that the selection and development phases were provided together, so that participants had the flexibility to pursue ideas in parallel. When not providing instructions the moderator was monitoring the teams as they completed the task. Monitoring the teams meant that participants could ask questions. The moderator would interject if the participants were significantly misinterpreting the instructions or discussing off-topic subjects and did not suggest particular usages of the CamBlend set-up.



1 Large postits
2 Mice (2 each side)
3 Keyboard
4 Clay and knifes
5 Whiteboard
6 Flip chart and pens
7 CamBlend Terminal

Figure 3. The lab layout for the task, including the seating numbers for participants used in the transcripts

Environment Setup

Participants were provided with a range of physical resources for the different task phases. Each participant was provided with a book of large post-its to generate their ideas, and a thick whiteboard pen. They were also provided with some clay and carving tools. As rapid prototyping is a common strategy for product design process, the clay allowed participants to produce a tangible version of their product for delivery. Figure 3 shows how these resources were configured for the participants.

Methods

To record the design session both the screens were recorded via screen capture software, and two fixed position camcorders were recorded each side of the interaction, making a total of 6 video feeds. One camera each side was the 'main' camera and one was backup in case of failure. This backup camera was useful either when participants turned their back to the camera, or the main camera missed some audio because of participant orientation. The video capture configuration resulted in 4 video feeds to analyse from each team (2 screen capture, 2 main camcorders). The 4 video streams were aggregated into a single video stream showing all the recordings together. This video consisted of a 2x2 grid of the source video streams, using the audio track of the video camera with the best audio quality. The analysis of this data used qualitative video analysis, including the orthography used in the presentation of the results [8], which draws on conversation and interaction analysis techniques [12]. Our primary analytic concern was our participants' practical accomplishment of co-orientation around objects, both local or remote, virtual or physical, as well as orientation around people.

CO-ORIENTATION

The following sections describe how co-orientation was achieved for various types of objects or people, and the main problems observed in achieving it. Initially we report generally on how groups approached the design task, starting with the first collaborative phase, idea presentation.

Presenting. Groups used all available methods, by either gesturing over the physical object, or directing using the focus window tools, or a combination of the two.

Clustering. All groups used this time to discuss their ideas in more detail. Those groups who did produce a 'clustering map' to structure their interaction, all opted to use the shared workspace provided by the system.

Selection. Some groups seemed clear on the idea they wanted to move forward immediately and therefore did so quickly, without much structure. Other groups were more formal in their selection, inventing complex voting methods that used the system features (highlight your favorite etc).

Development. A number of groups demonstrated developing several ideas in parallel before finally selecting one. Once development started groups seemed happy to split the workload into independent chunks, often conferring back to the main group at regular intervals with questions or updates. All teams used the modeling clay.

Referencing Objects

Participants from all groups seemed comfortable using deictic referencing to refer to physical objects while manipulating the focus windows of the system. Further, participants were able to manipulate the focus windows, whilst within a normal collaborative group discussion. So, for example, in the following fragment, the participants are presenting their ideas back to the group, and P2 is in the process of talking through his suggestions.

Team 6, fragment 14

```
2: red (.) glass, er, actual bottle (1.1)
3: What's this one? (0.3)
2: That one there's, that ones like I thought ..
```

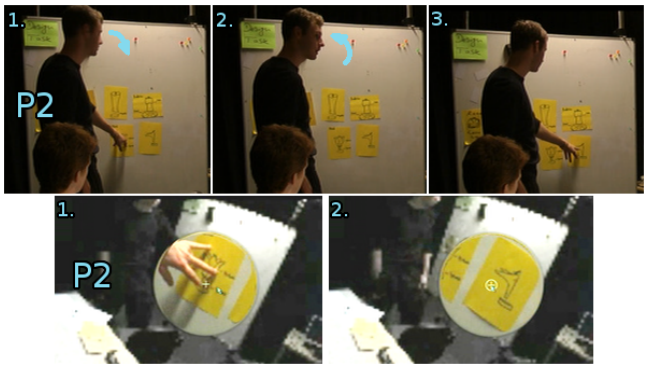


Figure 4. Top row: P2 glances at the focus window to resolve the point onto the new board item. Bottom row: The pointing motion acted by P3 (remote).

P2 is standing besides the whiteboard, describing the materials that will go into making idea 1, a flower. At the end of describing the flower, P3 (remote) interjects and asks about idea 2, a ship, adjacent to the flower. She does this by pointing at it by moving the focus window over from the flower to the ship. P2 then resolves the pointing action on screen to the ship. The resolution is done quickly, with only 0.3s between the end of the question and the beginning of the description. In order to answer the question, P2 must undertake some work to resolve what is being pointing at. To identify the item being referred to in talk, he must find the location of the focus window on the screen, then locate this back on the whiteboard. Here P2 was not observing the focus window movement at the point of its motion. This is important because no part of the pointing action seems to have been noticed by the remote participants as it was produced, the reference being resolved after the gesture was produced.

In this fragment a participant (P3) was successful in using the system tools to support the ongoing conversation. CamBlend allowed her to intervene into the flow of P2 presenting his ideas with a very specific, but seamlessly resolvable pointing action. Additionally, using the same resource and mechanism, this pointing motion could also have been to an item behind her. The system interface allowing the production of the pointing gesture was simple enough to allow this smooth conversational interjection. However, the fragment does highlight how the pointing production has changed. Using a screen based pointing system fragments the production phase of a gesture, which is revealed in this fragment. Instead of the face-to-face scenario where the production of a pointing gesture gains attention and forms the guidance towards its end-point [11], using the mouse splits the production into two pieces. Firstly, the physical mouse movement forms the foundation of the production. This action is invisible to the remote side for a variety of reasons, including the size, magnitude of movement required, and difficulty in translating into a meaningful action. Secondly the movement of the focus window itself can be observed, getting participant attention as it travels. This makes it difficult to follow a point in the same way as it might be done face-to-face, references are sometimes resolved after the gesture has been completed as they seem to be in this example. A feature of this type of screen pointing (and some other systems e.g. laser pointer based) is that it leaves a persistent record of the pointing action, which can be retrieved by the participant at a later date without requiring the producer of the gesture to in some way maintain the action until they see it is understood.

Team 2, fragment 13

A different type of physical referencing is shown in the following fragment where participants are selecting a final idea to report.

```
3: I was thinking [ this rom- this roman pillar,
1:                [ depends if
3: looked really good
1: Yeah put the shoe on that
```

Figure 5. Top row: The pointing motion which P3 controls. Bottom row: All participants are orientated around the screen at the point of confirming the reference.

During a conversation to move the focus window over to the roman pillar on the remote side, P3 interjects stating that it 'looks good'. P1 (remote) then confirms his understanding of which idea is being referred to by suggesting that it could be associated with a separate concept, the shoe. All four of the participants remain focused on the screen for the duration of the fragment.

In this example the pointing action didn't require remote participants to resolve the location of the object against the physical item behind them. The screen resources contained enough information for them to exclusively concentrate there. This meant that a key stage in the resolution of the pointing action was omitted, which was done at the expense of the ability to interact with or manipulate the object. Here it appears that the movement of the focus window was enough to draw attention to its position, causing the group to orientate around the contents of the window. The resolution of the object in question therefore relied on both their concentration on screen resources and monitoring of verbal deictic references.

Multiple Points of Entry

Participants quickly fell into using deictic references when coordinating action around local or remote physical or virtual objects. When a single person was using the system

and participants were focused on that action, this was quick to resolve. However, when multiple people were concurrently using the system resources, there were some problems as revealed in the following fragment.

Team 9, fragment 19

Here participants are clustering their ideas on the shared workspace. Before the fragment starts P1 has been tasked with recording the post-its from one physical board into the shared workspace, he does this in silence for the duration.

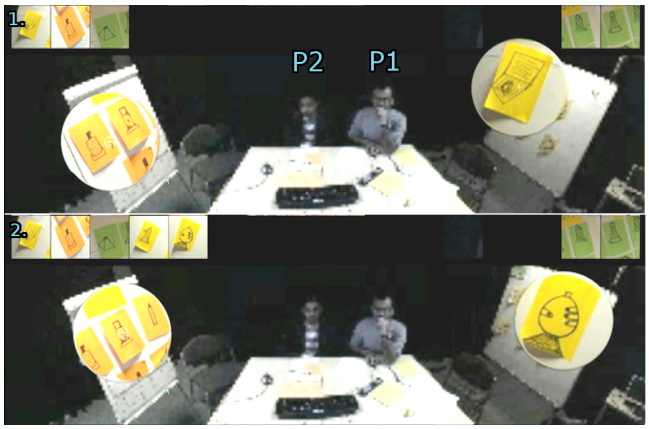


Figure 6. Remotely, P3 is manipulating the focus window on the left, while P1 is using the focus window on the right.

```
3: =Th- er, you can see the bubble yeah? (1.2)
2: E:r, [ yeah yep
3:      [ this one is the lilies one yeah?
(0.9)
2: E:r (0.6) which, yeah yeah, this one is
3: This one? (0.5)
2: Is the like simple one you know
3: These two are the simple ones?
2: Yeah
```

P3 (remote) here is requesting clarification of the ideas which P2 has created and put up on the whiteboard. P3 begins by positioning the focus window over one of his ideas and asks P2 whether he's able to 'see the bubble yeah'. Once confirmation is received he moves the focus window over the items, clarifying his understanding of each. He does this as P2 watches the position of the focus window offering feedback as it moves. At no point in this fragment does P2 look physically at the whiteboard he is sat near containing the physical post-its.

Here multiple people are using the pointing tools the system provides. As mentioned P1 is tasked with recording individual snapshots from the board in parallel to the activity outlined in the transcript. P3 (remote) seems aware of this as a point of confusion and uses two mechanisms to gain some common grounding with P2 before beginning the questioning. Firstly he explicitly mentions the focus window, as he is unable to relate his action around the specific focus window under control. He must therefore orientate around the screen resource by talking about 'the bubble'. Secondly he waits for confirmation before moving forward with questioning (1.2s).

Despite these two mechanisms P2 still takes sometime to find the correct focus window, over 1.5s considering the initial 'Er' is stalling while he resolves. Visually the cause of confusion is clear, with both focus windows in action participants have no way to resolve each against an owner. Both are in motion and therefore participants must use the context of the verbal questioning to deduce which of the focus windows is in question. Once participants gain this common ground the activity moves forward smoothly. With both participants orientating around the same focus window this affords a complex line of questioning where P3 moves and questions as P2 tracks and responds.

Putting Participants in Focus

In order to reflect face-to-face interaction, the technological features available were symmetrical for either co-locating around people or physical objects. They both occupy the same physical space and therefore the same tools are available to participants. However, the results show that these tools were used in distinctly different ways. Firstly the focus windows were rarely used explicitly as a mechanism to reference, handover, or exchange turns between participants. They had distinctly different uses and effects when orientating around people. The following episode is typical of the kind of reaction exhibited by participants.

Team 2, fragment 22

Here, team 2 are in the process of selecting one particular idea, discussing the merits of the different ideas presented.

```
2: Spend the next hour making funny faces on
(this)
4: ((laughs)) (3.7)
3: °Yeah I can't read it°=
2: =Quite disturbing how it's so zoomed in <lets
move it away>
```

Figure 7. P2 pulls as face, shown here as it is visible remotely

P3 (remote) is using the focus windows to look at the different ideas on the remote side and by chance leaves the focus window partially over the face of P2. After a pause, P2 pulls a face to the focus window, shown in figure 7, and moves the focus window off himself. The discomfort by P2 is clearly both shown and verbalized, he spends very little time sat with the focus window focused on him before feeling that it must be moved.

Generally speaking this behavior seemed to be related at least to the character of the person in focus, some participants seemed comfortable being zoomed in on. However the operation of the focus window had, more subtle consequences for the behaviour of participants. Pccaisionally participants were found to 'perform' to the focus window, adapting their behavior when the focus window was focused on them. The focus window seemed to

become a more explicit form of taking the stage, with some participants adapting their behavior in order to play to the focus window.

Team 9, fragment 23

In the following fragment from team 9, participants are in the process of deciding on an idea. P1 is asked to describe some detail of one particular idea from a participant on the remote side.

```
1: =I- [ It's like hands, like this, OK
3:     [ °Yeah°
1: as in giving [ your heart to a woman
3:              [ jus- just a (minute)
3: you [ 're not clear on our side
4:     [ OK
1: So it's like giving a ha- [ giving, like
4:                           [ OK
(0.2)
2: the concept [ is good
1:             [ it's like your hands gi- putting
1: your heart (.) in [ your hand, for a woman=
4:                   [ OK (OK)
```

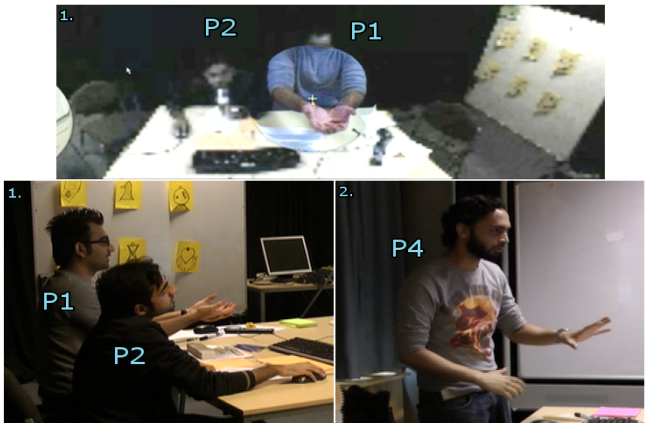


Figure 8. Top: The gesture as visible to the remote side. Left: The gesture produced physically. Right: P4 confirming his understanding after the last repeat.

Here P1 gives his description of the item a total of 4 times (the last time is not shown in the transcription), each time repeating both the same verbal description, and the presentation of his gesture. After the initial description, P3 moves the focus window over to his hands, prompting a second repeat. From here on P1 is literally in focus, and repeats his description including the presentation of his physical gesture (shown in figure 8) to support it 3 times more. Which finally results in P4 cutting him off in the middle of the final description.

The focus window here seems to be responsible for a form of 'taking the stage'. In the first instance the participants description seems casual, this description should have been sufficiently understandable to the remote side (other similar examples seem to have been). However, after the focus window is moved over to the participant his behavior changes distinctly. His description and action are more formal as he repeats himself several times. The decision to use the focus window here seemed to be disruptive to both participants. It was disruptive to P4 asking the question, by causing the repeated and unnecessary (demonstrated by P4's repeated confirmations) descriptions stopping the discussion moving on. It was also disruptive to the producer of the gesture, by causing the misinterpretation of the remote participants question and responses.

Team 6, fragment 49

In this fragment P2 is explaining a particular detail of his bottle design, which relates to the way the bottle twists as it extends.

```
2: er it's kind of like (.) it's kind of like I've
got er like a cylinder (0.7) and right let me
stand on the thing (0.8) And I've like (1.0)
2: <I've got the cylinder and I've like> squeezed
it (0.6) like that, like twisted it like that=
4:=Yeah, ah ok
```

Figure 9. Top: P2 re-orientates his gesture to fit the existing focus window. Bottom: The gesture as it appears remotely

P2 is asked about how the bottle design is twisted, which he would like to explain by producing a physical gesture to the remote side. He does this by physically re-orientating himself, leaning over P1 and lining his gesture up so that it directly overlays the focus window. He then pauses until the remote side are ready, then he produces the gesture, twice, after which he receives verbal confirmation. Here P2 seems to be monitoring the remote side to find when they're attentive to receive the physical gesture, pausing for 1.0s with his arms extended ready. The motion then seems to be quite simple to achieve. Little time is spent finding the focus window, or tweaking his hand position to be within it.

This fragment reveals how action is fractured between the screen and the physical spaces. Before the fragment P2 is working on the whiteboard, away from the mouse. In order for the remote side to recognise his gesture some work is required. He positions the focus window over himself by physically moving to the mouse and manipulating the system, and also positions himself over the focus window. P2 does seem to achieve this easily, demonstrating his sensitivity to the appropriate amount of visual information to produce and how the remote side are attentive to recognize it. This, however, highlights the requirements involved in resolving action within these two spaces.

CO-ORIENTATION IN CAMBLEND

Overall participants had few problems co-orientating around physical objects, virtual objects and people when using the CamBlend system. Where problems arose repair was consistently fast. However, the way co-orientation was achieved for different kinds of co-orientation was characteristically different, each revealing subtle problems.

Physical resources

There were various tactics for co-orientating around physical resources, local or remote. Participants could physically point or orientate around a resource when directing remote attention. Here, the position of the focus window on the screen allowed participants to remain sensitive to the remote domain. The focus windows themselves could be used, manipulating attention towards a physical resource as it appears on screen. Finally, verbal directives could be employed, describing the resource.

We found participants most often achieved co-orientation in either direction using the focus windows. On the whole they were successful using this tactic. However we found a number of limitations suggested in the fragments above. Sometimes participants found it difficult to orientate their action around individual focus windows (e.g. team 9, fragment 19). They had a limited means to differentiate their projective motion between any of the four windows, as the mouse pointer movement does not get transmitted and the physical movements of the mouse are not visible in the video. Occasionally therefore they relied on a number of tactics when referring to objects. As with other video-mediated techbologies they participants explicity mentioned the pointing device, focus window itself [15]. They moved the focus window in order to draw attention to it. To resolve the pointing action, participants needed to monitor the screen and verbal references to it in order to resolve what was being referred to. In other cases, participants would proceed to resolve the focus window position by referring to the physical world. The focus windows remaining on-screen persistently supported this. The person producing the gesture seemed to need to do less work to maintain the action and monitor fo completion of an activity.

Virtual resources

In co-orientating around virtual resources, participants did not have the option to refer physically or orient around a particular resource. Hence, they had a restricted range of resources available to them (i.e, the system pointing tools). Participants therefore faced a principal problem when co-orientating around virtual resources, relating their actions to specific screen resources.This was exacerbated by the number of resources available for the participants to orientate around. At the start of the task there were only 4 focus windows. As snapshots were recorded this increased the number of possible items to refer to. As described, participants occasionally found it difficult to discriminate amongst them. Further, some of the tactics they tried to use to refer to physical objects proved ineffective, for example it is not possible to wobble a snapshot as you can for a focus window. This might account for why we found that participants were generally reluctant to use the referencing tools in combination with deictic referencing. Often participants reverted to more explicit forms of reference, e.g. ‘the image on the left’.

People

The focus windows were very rarely used by participants to orient towards their colleagues. Co-orientation with other people was typically accomplished verbally by simply speaking in turn or addressing someone directly. Participants were therefore likely to exhibit the same more explicit forms of turn-taking described by O’Conaill & Whittaker [20], and the system tools are unlikely to have alleviated any of these co-orientation problems despite their success in referencing physical resources. When a focus window was directed towards a person, we observed some sensitivity to which part of the person was in focus. Hands and gestures were commonly used to convey meaning (e.g. Team 9 fragment 23 and Team 6 fragment 49). However, faces were rarely ever the focus, and when they were the participant in question seemed to feel uncomfortable (e.g. Team 2 fragment 22). This maybe because in Camblend it is very clear when someone is in focus, their face appears large and central to all four participants. Participants seemed to associate this focus with a stage, similar reactions to when one is performing, or when they are the focus of attention. This could result in repeated verbal descriptions and being more explicit in tone. Hence it seemed to be disruptive for both viewer and viewed. In face-to-face circumstances the differentiation between motions that support turn-taking and referencing objects is clear. Conversational turn taking contains a particularly complex array of social resources, including head-turning, glancing and facial expressions [23]. It does not generally include explicit pointing at someone who is speaking or about to speak.

FRACTURED ECOLOGIES

In everyday settings pointing, from its production, maintenance and retraction is highly complex in nature. However participants typically accomplish this and resolve the appropriate references unproklematically Luff et al. [15] describe how when mediated through video, an activity becomes fractured from the envionment in which it is embedded. The spatial inconsistencies that are introduced, when communicating through video, into asymmetries into how action is produced and how it is recognised. The problems introduced pervade all media spaces [6], but also CVE’s [10] and robot mediated collaboration systems [15].

CamBlend aimed to resolve many of these issues. By providing a panorama, most interactions do not lead to gestures that refer to items outside of the field of view. The pointing tools provided seem to allow users to recognize what pointing actions refer to. The focus windows by being placed in context help reconcile issues found in other systems [7, 25]. Virtual resources are integrated into the

same shared workspace through which participants communicate. Finally, providing both local and remote views through wide angle and symmetrical views of both spaces, mean that viewpoints, occlusion and line of sight are consistent. However it seems that CamBlend factures conduct in different ways.

Relating Screen to Local Environment of Action

First, using a screen-based representation of the interaction space, seems to fragment this screen resource, and the local environment in which participants's actions are embedded. This issue was only hinted at in a previous study of CamBlend [19]. Referencing physical resources through their screen-based representations seemed to involve participants engaging in additional activities in order to resolve the on-screen reference with the referenced object. Even in cases where the interaction seemed seamless, (e.g. team 6, fragment 14) there was a perceptible delay in doing this work.

Making local video available on-screen is a common strategy, e.g. Skype [1] implements a video feedback window as inset to the view of the remote space. The video view of the local environment can also be used as a shared space, creating a common resource between two parties but requiring one side to resolve the video feed to the physical space, as in [4]. This shared view can also be annotated by the remote helper, as in [5], supporting the worker to resolve the instructions drawn over the video to the model in front of them. Apart from this video augmentation, CamBlend does not provide any additional tools to help resolve the two spaces. Neverthelessm participants seemed quite accomplished at doing so.

Arguably, a combination of the task design, with the unification of virtual and shared physical resources allowed participants to rely on the video feed as an informational resource and sacrifice the ability to manipulate the object. It was surprising the extent to which participants were able to ignore their local physical surroundings (e.g. team 2 fragment 13) when undertaking the activities reported on here. The fragmenting of action did not seem to cause severe interactional disruption to the participants as all the resources were integrated on-screen. In those situations where people needed to refer to the physical space, participants were able to choose to switch between the on-screen representation of the local space and actually looking at the physical space. Participants switched as infrequently as possible, seeming to be aware of the appropriate amount of information they need to provide at any moment.

Relating to Multiple Screen Resources

The complexities that arose when using the available system tools to orient to an object seemed to fracture conduct in a second, different way, Normally when producing a pointing action participants need to tie their talk with the movement of the focus window. We found that because participants could not orientate themselves around the various different screen resources, this inhibited their ability to tie the verbal directive with any one of the particular resources those within the environment, the physical and the digital as well as the images of people on screen.

The need to relate to or reference objects on-screen is common in collaborative systems. Most video communication tools support the display of and interaction with shared resources (e.g. through screen sharing) and these resources tend to appear next to the video. In contrast, CVEs integrate users and digital resources in the same virtual interaction space [10]. In both cases, interaction can be referred back to a specific user, as every user has typically one pointer. Even in these cases with only one pointing device such as in the GestureMan experiments [15] additional work is necessary to resolve a pointing action, particularly with respect to how participants project future conduct. CamBlend provides additional resources for resolving references. Firstly it integrates virtual resources with the physical pointers on a single screen. Secondly it provides 4 pointers with the aim that multiple participants have the capability to point simultaneously to physical and virtual resources.

The study reported here found that this additional functionality fragmented the possible focus of any verbal directive. Participants compensated for this by reverting to mechanisms to establish a common referent before continuing, including wobbling the focus window and verbally talking about the focus window. These available resources were limited though, and so occasionally caused a breakdown in conversation as both participants needed to resolve which item was of interest.

CONCLUSION

Broadly speaking, fractured interactional ecologies in video communication, described in previous work and in this paper result from two principal but conflicting aims. System designers aim to support naturalistic interaction embedded in physical space, around physical resources and people. Ideally, no other resource but the view would be required. We know however, that the limitations of the camera pin-hole model and available forms of display severely limit the ways in which this can be achieved. To overcome those limitations but also to actively support digital interaction around physical and digital resources from multiple places, a variety of additional resources need to be hen introduced, these include additional views, pointing tools and digital representations of physical resources, each of which are in one way or other removed from physical space itself. To make accomplish actions in interaction, participants need to do translate between what is available in the interaction space and what is represented of it on screen.

The CamBlend system creates one way of addressing these these two objectives. It provides a distinctive way for participants to interact around actual physical resources whilst allowing interaction around digital screen resources.

The system thus allows us to closely examine one particular kind of problem – a problem that is critical to address when developing systems to support collaboration – that of reference. For range of technological solutions, particularly ones which aim to provide a rich environment for interaction this problem has been particularly hard to resolve when participants engage in remote activities. Here, we have focused on one way of trying to provide an environment in common where participantts can reach in to a remote domain and engage with others in activities with objects [5, 15]. Like others we have found there are problems to resolve, particularly given the fragmented domains in which the participats need to work. However, through such technical prototypes and interventions we can begin to layout the ways in which we might betyter design technologies that allow participants to produce coherent material conduct in virtual space.

ACKNOWLEDGMENTS

We thank all study participants and gratefully acknowledge support from the EPSRC through grants EP/P504252/1 and EP/J006688/1, the Creativity Greenhouse project.

REFERENCES

1. Skype. http://www.skype.com/
2. Aanestad, M., The Camera as an Actor Design-in-Use of Telemedicine Infrastructure in Surgery, Computer Supported Cooperative Work, 2003, 1-20
3. Baudisch, P., Good, N., Stewart, P., Focus Plus Context Screens: Combining Display Technology with Visualization Techniques., Proc. UIST, 31-40
4. Fussell, S.R., Kraut , R.E., and Siegel, J., Coordination of communication: effects of shared visual context on collaborative work, in Proc. CSCW'00 ACM (2000)
5. Fussell, S. R., Setlock, L.D., Yang, J., Ou, J., Mauer, E. M., & Kramer, A., Gestures over video streams to support remote collaboration on physical tasks. Human-Computer Interaction 2004, 273-309
6. Gaver, W., The affordances of Media Spaces for Collaboration., Proc. CSCW '92., 17-24
7. Gaver, W.W., Sellen, A., Heath, C and Luff, P., One is not enough: multiple views in a media space, in Proc. INTERACT '93 and CHI '93. ACM (1993), 335-341.
8. Heath, C., Hindmarsh, J., and Luff, P., Video in Qualitative Research, SAGE.
9. Heath, C., Luff, P., Kuzuoka, H., Yamazaki, K., Oyama, S., Creating Coherent Environments for Collaboration, Proc. CSCW '01, 119-138
10. Hindmarsh, J., Fraser, M., Heath, C., Benford, S and Greenhalgh, C, Object-focused interaction in collaborative virtual environments. TOCHI '00, ACM (2000), 477-509
11. Hindmarsh, J. and Heath, C., Embodied Reference: A Study of Deixis in Workplace Interaction. Journal of Pragmatics, 1999. 32: p. 1855-1878.
12. Jordan, B and Henderson, A., Interaction Analysis: Foundations and Practice. in Journal of Learning Sciences, PARC (1994), 39-103
13. Kuzuoka, H., Yamazaki, K., Yamazaki, A., Kosaka, J., Suga, Y., Heath, C., Dual ecologies of robot as communication media: thoughts on coordinating orientations and projectability. Proc. CHI'04, ACM Press (2004), 183-190.
14. Kuzuoka, H., Kosaka, J., Yamazaki, K., Suga, Y., Yamazaki, A., Luff, P., Heath, C., Mediating dual ecologies, Proc. CSCW 2004., 179-182
15. Luff, P., Heath, C., Kuzuoka, H., Hindmarsh, J., Yamazaki, K and Oyama, S, Fractured Ecologies: Creating Environments for Collaboration. In Human-Computer Interaction 2003, 51-84.
16. Luff, P., Yamashita, N., Kuzuoka, H and Heath, C, Hands on Hitchcock: embodied reference to a moving scene, In Proc. CHI '11, ACM (2011), 43-52
17. Martin, D., Rouncefield, M., Making the organization come alive: Talking through and about the technology in remote banking., Human-Computer Interaction, 2003, 111-148
18. Nardi, B, A., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S., Sclabassi, R., Turning away from talking heads: the use of video-as-data in neurosurgery. Proc. CHI'93, ACM Press (1993), 327-334.
19. Norris, J., Schnädelbach, H., Qiu, G., CamBlend: an object focused collaboration tool. Proc. CHI'12, ACM Press (2012), 627-636.
20. O'Connaill, B., Wittaker, S., Characterizing, predicting, and measuring video-mediated communication: A conversational approach., Video-mediated communication, Erlbaum, 1997, 107-132
21. O'Hara, K., Kjeldskov, J., and Paay, J., Blended interaction spaces for distributed team collaboration. In TOCHI '11. ACM (2011)
22. Osborn, A, F., Applied Imagination, Scribner 1963
23. Sacks, H., Schegloff, E.A and Jefferson, G. A simplest systematics for the organisation of turn-taking for conversation. Language, 1974, 696-735
24. Streeck, J. On Projection. in Interaction and Social Intelligence (ed by E. Goody), Cambridge University Press, Cambridge, 1995, 84--110
25. Yamaashi, K., Cooperstock, J, R., Narine, T., Buxton, W., Beating the limitations of camera-monitor mediated telepresence with extra eyes, in Proc. CHI '06. ACM (1996), 50-5

The columns on the last page should be of approximately equal length.
Remove these two lines from your final version.